Testing Big Data Application

Conference Paper · September 2019

DOI: 10.23919/EECSI48112.2019.8976972

4 authors, including:

Narinder Singh Punn
Indian Institute of Information Technology Allahabad
49 PUBLICATIONS **1,078** CITATIONS

SEE PROFILE

Sonali Agarwal
Indian Institute of Information Technology Allahabad
195 PUBLICATIONS **2,959** CITATIONS

SEE PROFILE

Testing Big Data Application

Narinder Singh Punn
Dept. of Information Technology
Indian Institute of Information Technology Allahabad
Prayagraj, India
pse2017002@iiita.ac.in

Sonali Agarwal
Dept. of Information Technology
Indian Institute of Information Technology Allahabad
Prayagraj, India
sonali@iiita.ac.in

M. Syafrullah
Program of Master of Computer Science
Universitas Budi Luhur
Indonesia
mohammad.syafrullah@budiluhur.ac.id

Krisna Adiyarta
Program of Master of Computer Science
Universitas Budi Luhur
Indonesia
krisna.adiyarta@gmail.com

***Abstract*—Today big data has become the basis of discussion for the organizations. The big task associated with big data stream is coping with its various challenges and performing the appropriate testing for the optimal analysis of the data which may benefit the processing of various activities, especially from a business perspective. Big data term follows the massive volume of data, (might be in units of petabytes or exabytes) exceeding the processing and analytical capacity of the conventional systems and thereby raising the need for analyzing and testing the big data before applications can be put into use. Testing such huge data coming from the various number of sources like the internet, smartphones, audios, videos, media, etc. is a challenge itself. The most favourable solution to test big data follows the automated/programmed approach. This paper outlines the big data characteristics, and various challenges associated with it followed by the approach, strategy, and proposed framework for testing big data applications.**

Keywords—Big data, Testing, MapReduce, Hadoop.

I. INTRODUCTION

Generally, with the term Big Data, we get the idea of a huge volume of information. Big data computation frames another pattern for inescapable processing with the degree of information developing and the quickness of data expanding. With the origin of new advancements, an enormous amount of organized and unstructured data is delivered, gathered from different sources like social media, audios, websites, video and so forth which is hard to oversee and process [1]. The following figure 1 shows various sources of stream data. With this wide variety of data sources, it poses challenges to the testing community.

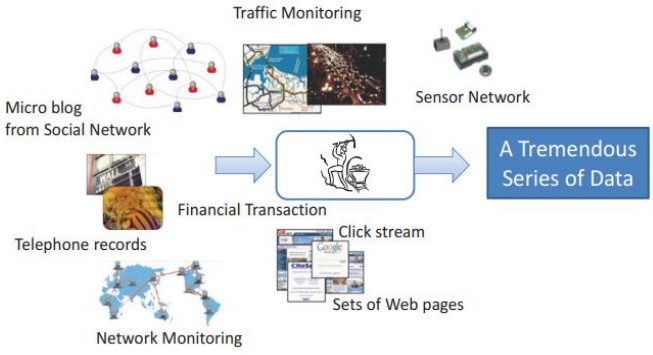

Fig. 1. Data Sources

A. *Why Big Data Testing?*

Big data possess the characteristic property of volume, velocity, variety, and veracity, which makes its behavior dynamic in nature. And this nature of big data streams makes the testing process crucial, which if not performed efficiently will pose adverse effects on organizations. Big data testing benefits the organizations substantially by providing data accuracy, better decision making; which plays a vital role in any business, better market strategy, reduces the deficit and boost profits and a lot more. For testing big data, it requires that the tester must have an optimal level of understanding what big data framework is followed. Today, with the huge volume of data generated, QA expertise can have a hard time managing them. And this makes it quite crucial to perform big data testing and harness its advantages.

B. *Big Data Characteristics and Data Formats*

Though there are many V's introduced, following 4Vs define the big data characteristics. These are *Variety, Velocity, Volume* and *Veracity* as shown in figure 2 [2, 3].

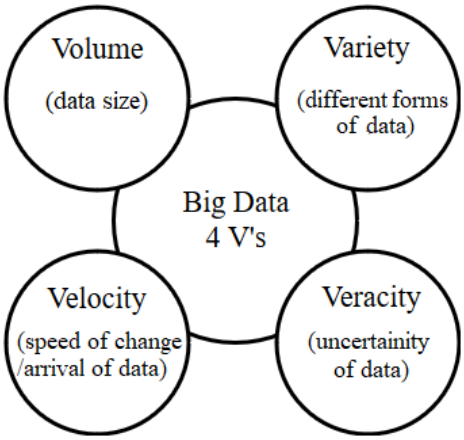

Fig. 2. 4 V's of big data

Variety: (different forms of data) Data or information that comes in for the processing can be of a variety of forms and formats. For instance data can come in different file formats like .txt, .csv, .xlsx, etc. Sometimes the information may not be formatted in the desired manner e.g.; data can come in the form of SMS, audio, video, pdf or other doc format or something we may not have contemplated it. It makes it quite crucial for the organizations to handle such a wide variety of data efficiently as at present wide range of formats of data are available to seek information from it.

Velocity: (speed of change/arrival of data) This characteristic of big data provides the glimpse of the pace of data i.e., at what rate data is arriving in from various sources like networks, social media, and other business processes. This high-speed real-time data is massive and comes in a

continuous fashion which may need immediate processing. There is even possibility of mutation in the data over time.

Volume: (data size) At present data comes in/generated from different sources by machines, networks, and media, from which valid information is extracted and aggregated at the organization hub. For instance consider the case of a variety of data sent by smart mobile phones to the network infrastructure, the information collected from various surveys, feedback forms, etc., this aggregated information forms the enormous size of data which needs to be properly analyzed.

Veracity: (uncertainty of data) There are a wide variety of sources of data stream available which produce a huge amount of data. With these many available sources, this data becomes vulnerable to outliers or noise. Due to which the nature or behavior of the data may change. The term Veracity describes this as the uncertainty of data which poses a huge impact on the decision-making process of the organization.

Based on the above characteristics, data comes in with different sizes, formats, rate, etc., thereby resulting in the following categories of data formats [3]: *Structured, Unstructured, and Semi-structured* as shown in figure 3.

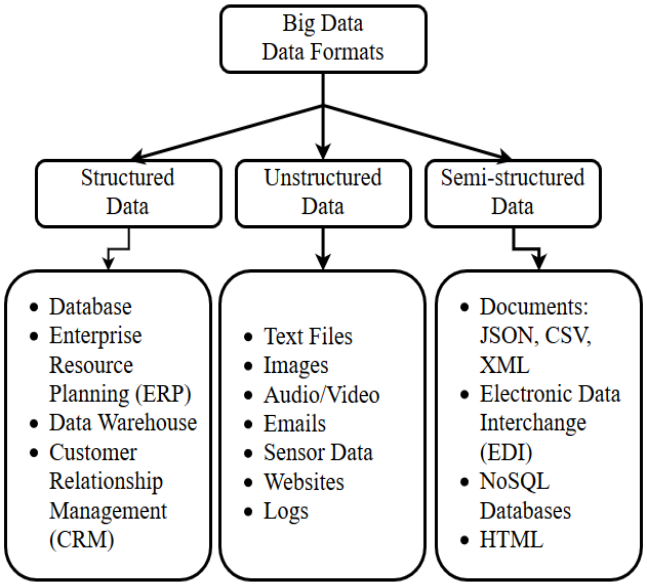

Fig. 3. Big Data: data formats

Structured Data: (high degree of organization) Structured data comprises of definite design in a well-organized manner such that data is easily utilized for processing and analysis. As an example, we can consider the relational database structure in which information is stored with some standards such that the same standards can be utilized in retrieving the information. This data format has a relational key and can be easily mapped into pre-designed fields.

Unstructured Data: (low degree of organization) This type of data does not follow any kind of pattern. This makes it quite difficult to analyze and may require some tools to extract desired information for processing, for example, streaming data, web pages, electronic mails, videos, etc. This lack of organization of data makes it a time-consuming task to process data. Most of the real-time data stream is unstructured. However, by identifying some hidden pattern, we can utilize it for the desired purpose but finding that pattern is difficult and time-consuming.

Semi-Structured Data: (partially organized data) This type of data can be organized by applying a bit of desired operations as in conversions, shifting, etc. There is various software that handles this kind of data like Apache Hadoop. Sometimes we call this type of data as structured data, which is lacking some sequence or pattern and is available in an unorganized manner. Such kind of information can come in the form of tab-delimited text files, CSV files, BibTex files, XML, web data such as JSON (JavaScript Object Notation) files, and other markup languages. This type of data is easier to handle as compared to completely unstructured data where we need to spend little effort to identify the pattern and process the information.

II. Related Work

With the ever-increasing essence of data, its storage, security, visualization, analysis, and testing have become the challenging task, in regard with this dynamic nature of big data, there are several works proposed from various areas of research.

Galit Shmueli et-al. [4], introduced a super-power large-sample approach for analyzing the application hypothesis with huge samples and also focused on the inadequacy of employing the small-sample approach for testing considerations with these large samples. Due to the reduced sampling and high statistical power of large samples can be utilized for hypotheses testing. The super-power methodology consolidates the whole data examination process, through large data samples, then determining conclusions. Mustafa Batterywala et-al. [5] highlights the performance testing of big data applications to tackle challenges in outlining testing strategies, tools, test environment, etc., by considering data ingestion, throughput and data processing capacity, map-reduce operations, and so on as a primary candidate for performance testing. To obtain optimal performance from the big data environment requires continuous monitoring in critical areas of data storage, commit logs, thread concurrency, and timeouts, caching, JVM parameters, map-reduce operations and message queues. To overcome the quality challenges in big data, Mahesh Gudipati et-al. [6] introduced a functional and non-functional testing approach. They correlated best testing techniques and took after most useful modes to enhance the testing quality, which helps in identifying the deformations at the early stage and reduce the risk of failures.

Testing of big data applications can be made efficient by designing and executing test plans; approach and strategy for all V's of big data [2]. Furthermore, testing techniques can be introduced for testing of response time, user interface, user requirements, etc. Hadoop offers an excellent platform to test the big data applications [7] since it supports structured, semi-structured, and unstructured data formats in the distributed environment and moreover it's open-source software.

III. Big Data Testing Approach And Strategy

Big data testing [8] is often associated with varied sorts of testing, for example, functional testing, performance testing, database testing, and infrastructure testing. Along with these, it is critical to have an unmistakable test plan that permits a simple rendering of big data testing. When performing big data testing, comprehend that the idea is mainly about checking the application capability to handle thousands of gigabytes of data. Big data testing for CPS are

often generally partitioned into three vital stages that incorporate:

Data staging validation: Also referred to as a pre-Hadoop stage, the method of big data testing starts with process validation, which aids in guaranteeing if right data is pushed into the "Hadoop Distributed File System (HDFS)". Validation testing is done for the data which is taken from different references, for example, RDBMS, webisodes, and online networking. Then the data is coordinated with the data utilized in the Hadoop process so as to check if the two coordinate with one another. A number of normal tools that might be utilized for this stage are "Talend and Datameer".

MapReduce validation: It is the idea of programming that takes into account colossal adaptability crosswise over a huge number of servers in a Hadoop cluster.

Amid big data testing, MapReduce second stage is the validation in which a tester analyzes the legitimacy of business logic on every joint which is taken post the validation of the previous data after running opposite various joints. This aide in guaranteeing that:

- The procedure of MapReduce is functioning without any flaws.
- Date accumulation or isolation rules are accurately performed on the data.
- Value key sets are produced properly.
- After the MapReduce process, data validation takes place.

Output Validation: On effectively performing the initial two stages, the last stage of the method is "output validation". It incorporates processed files which are prepared to be passed to an "Enterprise Data Warehouse (EDW)" or some other system in the view of particular necessities. Output validation stage incorporates the below-mentioned steps:

- Need to validate change rules are effectively applied.
- Need to validate the data respectability and in addition effective data lading into the resulting organization.
- Guaranteeing that any data defilement by differentiating the objective data and the HDFS file organization data.

A. Performance Testing

Performance Testing [5] for Big Data incorporates the following activities:

Data ingestion and Throughout: During this activity, the tester checks for the speediness of the system in which it can consume data from numerous data references. Testing incorporates the recognizing of different messages that can be processed by the queue in a given time. It additionally incorporates how rapidly data can be implanted into the underlying data store for instance insertion rate into a Mongo and Cassandra database.

Data Processing: It incorporates the conformance of the speed with which the queries or map-reduce jobs are performed. It likewise incorporates testing of the data handling in segregation when the repressed data store dwells inside the data sets. For instance, running Map Reduce jobs on the underlying HDFS.

Sub-Component Performance: These are the systems which are comprised of various parts, and it is required to validate each of these parts in separation. As per given example which shows the rapidness with which message is listed and devoured, map-reduce jobs, query performance, search, and so on.

B. ETL Testing

ETL is an acronym for extract, transform and load, and has been around for a quite a while in light of the fact that it is related with conventional batch processing in the data domain [9]. The purpose of data warehouses is to give businesses with data that they can solidify, analyze, and make cognizant thoughts out of that is pertinent to their focus/intents. There are ETL tools through which the crude data is changed over into a meaningful format. The tool likewise aides them changes over data into a format that can be utilized by businesses. Software merchants like IBM, Pervasive, and Pentaho provide ETL software tools.

Extract: Once the data is gathered, it will be extricated/ perused from the source database. This is done to every one of the databases.

Transform: Transformation of the data is done straight-forward. The data format is changed into usable chunks and must conform to the requirements of the target database.

Load: This is the last stage where data is written to the target database.

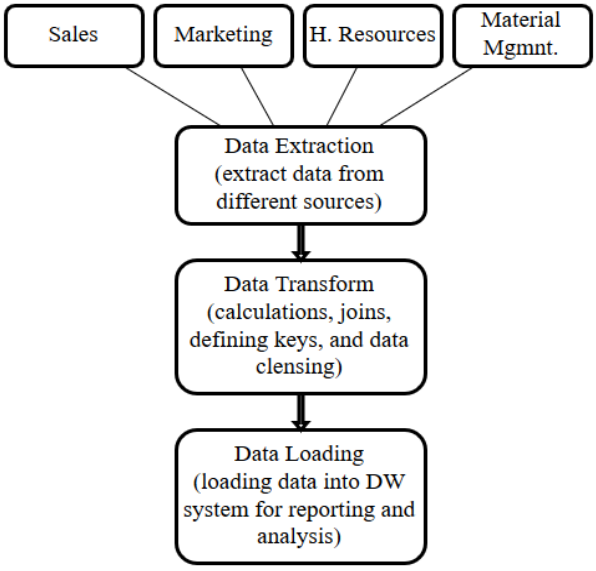

Fig. 4. ETL Process

To guarantee that the data secured in this way is reliable, tools for data integration processes like data profiling, cleansing, and auditing are all merged with data quality tools. This whole procedure as shown in figure 4 will guarantee that you have extracted actual data. ETL tools are likewise crucial for loading and converting both structured and unstructured data into Hadoop. It additionally relies on the kind of ETL tools that you use. Highly advanced ones let you convert multiple simultaneously.

The data processing part in a data warehouse follows a three-layer architecture during the ETL process.

Data Warehouse Staging Layer: The staging area is a transitory area or a landing zone where data from every one

of the assets are stored. This zone guarantees that all the data is accessible before it is coordinated into a data warehouse. It is basic for the data to be placed somewhere as a result of fluctuating business cycles, hardware impediments, network resource restrictions, and data processing cycles. You cannot extricate all the data from all the databases in the meantime. Consequently, data in the data warehouse is transient.

Data Integration Layer: This is the establishment of next-generation analytics and it contributes to business insight. The data integration layer is a blend of semantic, reporting, and analytical technologies based on the semantic knowledge framework. Data is arranged in hierarchical groups known as facts and changed over into aggregated facts. This layer is the connection between the staging layer and the database.

Access Layer: Using common business terms, users will be able to access the data from the warehouse. The access layer is what the users can get to, and the users themselves know what to make of the data. It is relatively like a virtual layer since it doesn't store information. The layer contains data-focused to a targeted population, making access and usage simpler.

C. Other Testing Approaches:

Regression Testing: This testing technique centres around the errors that may occur due to the client's improvement request, bug fix, new feature addition, or after doing any changes in the application [10]. This can be carried out by the following techniques:

Reset all: This is one of the methods for regression testing in which all the tests in the existing test bucket or suite should be re-executed. Since this can be very expensive on a standalone system, Hadoop serves as a promising platform to execute these test cases parallel in the distributed environment. *Regression test selection:* This considers part of the test suite (reusable test cases and obsolete test cases) to be run. *Prioritization of test cases:* Prioritize the test cases depending on business impact, critical and frequently used functionalities. Selection of test cases based on priority will greatly reduce the regression suite.

Failover Testing: The effort highlights on favouring the procedure for recovery due to any kind of failure. Failover testing is done to analyze the system's capability to guarantee data recovery, preventing data corruption, managing edit logs by introducing critical failures or exploiting the performance thresholds [6]. Recovery Point Objective (RPO) and Recovery Time Objective (RTO) are the metrics that can be utilized to quantize the failover test suite.

IV. Future Scope

Based on the pace with which data is increasing, there is a huge need for research in the area of big data testing. There are several other testing techniques that can be explored like a genetic algorithm, particle swarm optimization, and so on. Big data testing techniques can further be improved with the help of data exploration process that cites to obtaining the comprehension of the dataset with the goal of producing hypotheses, testing assumptions, supporting the selection of statistical methods and providing a premise for further data accumulation. Advancements can be followed to have a better and easy understanding of the various challenges and issues pertaining to big data, and some new factors that may arise because of its dynamic nature.

Indeed there is a lot to be researched on big data testing as data is evolving every single day and a new type of challenges are originating which are leading towards new research fields corresponding to it.

V. Conclusion

Big data testing is quite different from regular software assessment and plays a vital role to cope with various corresponding challenges of it. The best approach for testing big data applications is to incorporate data quality maintenance, data sampling technique, and automation of test suite. Since data is present in various formats which can be structured, semi-structured or unstructured, it requires some pre-processing. With the help of some tools and techniques, we can format the unstructured data into semi-structured data, and even further processing can make this semi-structured to structured data. The dynamic nature of big data can introduce new challenges and exploring these new challenges and examining various other tools and techniques can be considered as a future scope or extension of this paper.

References

[1] Chen, Min, Shiwen Mao, and Yunhao Liu. "Big data: A survey." Mobile networks and applications 19.2 (2014): 171-209.

[2] Adiba Abidin, Divya Lal, Naveen Garg, Vikas Deep "Comparative Analysis on Techniques for Big Data Testing," 2016 InCITe

[3] Zikopoulos, Paul, and Chris Eaton. Understanding big data: Analytics for enterprise class hadoop and streaming data. McGraw-Hill Osborne Media, 2011.

[4] Shumeli, Galit, Mohit Dayal, and Bhimasankaram Pochiraju. "Testing Theories with Big Data: A Super-Power Approach." (2012).

[5] Batterywala, Mustafa, and Shirish Bhale. "Performance Testing of Big Data Applications." Impetus Technologies, STC (2013).

[6] Mahesh Gudipati, Shanthi Rao, Naju D. Mohan and Naveen Kumar Gajja, ³Big Data: Testing Approach to Overcome Quality Challenges´, Infosys Labs Briefings, Vol 11, No 1, 2013

[7] Campos, Jaime, Pankaj Sharma, Unai Gorostegui Gabiria, Erkki Jantunen, and David Baglee. "A big data analytical architecture for the Asset Management." *Procedia CIRP* 64 (2017): 369-374.

[8] H. M. Sneed, K. Erdoes, "Testing big data (Assuring the quality of large databases)", 2015 IEEE Eighth International Conference on Software Testing Verification and Validation Workshops (ICSTW), pp. 1-6, 2015.

[9] Mukherjee, Rajendrani, and Pragma Kar. "A comparative review of data warehousing ETL tools with new trends and industry insight." In 2017 IEEE 7th International Advance Computing Conference (IACC), pp. 943-948. IEEE, 2017.

[10] Alexandrov, Alexander, Christoph Brücke, and Volker Markl. "Issues in big data testing and benchmarking." In *Proceedings of the Sixth International Workshop on Testing Database Systems*, p. 1. ACM, 2013.